Cut to Fit: Tailoring the Partitioning to the Computation
Iacovos G. Kolokasis, Polyvios Pratikakis
Foundation for Research and Technology Hellas (FORTH)
Computer Science Department, University of Crete

Graph Processing
- Social Network Analytics computations are a significant part of big data applications
- Data placement strategy affects the performance of the analytic framework

Partitioning and Placement
- Investigate how knowledge of the application and the dataset can help optimize the performance with minimal effort
- We concentrate on the impact of the partitioning strategies on the performance on computations on social graphs
- We introduce two new partition strategies: Source Cut and Destination Cut
- Provide a smart agent to select – not the best for all – an efficient partition strategy for a given graph algorithm

Dataset Analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Vertices</th>
<th>Edges</th>
<th>Symmetry</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoadNet-CA</td>
<td>1.9M</td>
<td>5.5M</td>
<td>100.0</td>
<td>Low-Degree</td>
<td>83.7MB</td>
</tr>
<tr>
<td>socLiveJournal</td>
<td>4.8M</td>
<td>68.9M</td>
<td>75.03</td>
<td>Heavy-Tailed</td>
<td>1.0GB</td>
</tr>
<tr>
<td>follow-dec</td>
<td>26.3M</td>
<td>204.9M</td>
<td>37.57</td>
<td>Power-Law</td>
<td>4.1GB</td>
</tr>
</tbody>
</table>

Characterization of datasets.

In-degree and Out-degree distribution of graph datasets.

Graph Partitioning

Partition Metrics

| Replication Factor | Balance Partitions |

Smart Agent
- We run experiments using various type of datasets, with 128 and 256 number of partitions
- Through the experiments, we have found that in general case the two most efficient partitioners are 2D and DC
- We tested various heuristics to achieve the best fit according the results
- Heuristic select the partitioning granularity based on the dataset size and the number of partitions

Evaluation

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Instance</th>
<th>Cores</th>
<th>Memory</th>
<th>CPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master</td>
<td>1</td>
<td>32</td>
<td>128GB</td>
<td>Intel(R) Xeon(R) E5-2630 CPUs</td>
</tr>
<tr>
<td>Workers</td>
<td>4</td>
<td>32</td>
<td>256GB</td>
<td>Intel(R) Xeon(R) E5-2630 CPUs</td>
</tr>
</tbody>
</table>

Conclusions
- Graph analytics computation are dependent on the properties of each graph, the application needs and the number of partitions
- Replication factor not affect the performance in all cases
- Provide Smart Agent, a heuristic to select the partitioning granularity based on the dataset size and the number of partitions
- Smart Agent achieve the second best performance execution
- Smart Agent underperforms from the best execution time at Pagerank 12%, Connected Components 2.7%, Triangles 2.50% and Single Source Shortest Path 2.6%

Acknowledgement
We thankfully acknowledge the support of the European Commission under the Horizon 2020 Framework Programme for Research and Innovation through the EUROEXA (H2020–GA–754337) project.